Problem Statement Definition

If you designed your own problem statement (X: Design Your Problem Statement), please provide details on it here.

*Your problem statement should cover:  
Identify the problem you are trying to solve & What is the challenge? (https://graphforall.devpost.com/details/inspiration#h\_79624773520441643228989294)*

GRAPH FOR BETTER KNOWLEDGE LINKING

X: Next Generation Knowledge management to showcase personal decision support for maintaining a sustainable life style

Author: Dr. Mirko Kämpf & Eric Tessenow

Introduction:

With the Green Deal in Europe and other initiatives to integrate sustainability in our lifestyle, we often struggle with how we can be more sustainable and even more if our surroundings are allowing us to be.

What is the problem, or class of problems?

One of the choices we make is the place where we live, i.e., when we buy a property, we are looking at a lot of factors that can guide us whether the surroundings fit our needs. Unfortunately, not every relevant aspect is present in an expose nor is the data accessible easily.

We have to combine lots of different knowledge bases to create a more holistic understanding before we are able to apply interesting analysis algorithms which then help us to find answers to important questions which matter in real life.

The field of data science and data engineering has been growing in the last two decades from single computer applications to scalable cloud-based solutions. Special purpose databases and analytics engines appeared. Those solve many classes of important problems, such as data lake management on scalable clusters with storage, and compute capabilities, real-time streaming data integration, API based system integration, just to name the most obvious.

Technical standardization is ongoing, and unified tool support many projects on multiple levels, i.e., the data import and export across data centers and access to streaming data across continents are well supported by robust open protocols and enterprise ready data management solutions. Reliable exchange of files, messages, events, or database records is not a big problem any longer, due to the progress in recent technologies on which we can build our systems upon.

But the integration of data models from multiple domains is still a non-trivial task.

Ontologies have been developed to express the knowledge models of existing systems,

Knowledge graphs are in the focus of many projects, products, and start-ups.

But how can we combine all those knowledge spread across many graphs in such a way that allows us for a flexible information management and data analysis, using ML and AI systems, without becoming an export in all the related domains?

Or in short: How can we link data from multiple sources very efficiently to prepare the input for advanced spatial analysis, combined with time-series analysis, structural analysis, and NLP?

What is the challenge?

Many well specialized data analysis platforms do exist. And most of them have their special optimized data structures and data representations, which typically match their goals best.

The overarching data model we consider is a property graph. Using nodes and edges with specific attributes allows us to represent a variety of data structures, managed in "special purpose data management solutions". Staging the data in a scalable graph database can be a solution which simplifies the information integration processes across organizations, enterprises, and individuals.

What particular problem do we address?

In order to get a better understanding of the surroundings of a property or the place we visit on a trip we might be interested in how sustainable life in this location could be. We know the projects OpenStreetmap, and Wikidata. In theory, we could start digging in the public data using cloud based systems.

But this is far away from convenient information retrieval and requires a lot of expertise in the OSM and Wikidata data domains. It rather is a data science and data engineering exercise. It includes using multiple knowledge sources which must be combined technically, and logically so that the contained facts are related to each other in a reliable and meaningful way.

Assume a simple and reusable load operation for the graph database, which maps the raw data into a knowledge model. With this we can establish scalable integrated graph layers, to represent a variety of reusable data models which all fit particular use cases.

We would be interested in investigating individual aspect of, let’s say, sustainability related facts in a particular layer on its own, or any combinations of them. And by adding variable personal preferences to the data structures, we would be able to feed a powerful analytics engine which helps us to understand the interaction of several aspects, represented as graph layers, and to cater to the needs of an individual, a community, or the society. The most important aspect of such an approach would be it’s ability to provide personalized results, rather than just generic maps which are sometimes hard to interprete.

Such a model could help in the future local communities better define needs to become more sustainable. It could become a good decision support for individuals to where they want to settle down or offer the missing services, needed by a sustainable community.

Resources:

- https://github.com/GeoQB/geoqb

- https://journals.plos.org/plosone/article/authors?id=10.1371/journal.pone.0141892

- https://www.openstreetmap.org/

- https://www.wikidata.org/wiki/Wikidata:Main\_Page

- <https://www.tigergraph.com/>

\* Please explain the following (Max 250 words) for Impactfulness:

***How does your project positively impact people’s lives?*** *In your response, address the following: How many people, families and organizations does your solution help? What’s the level of social and/or economic benefit to the world and how does your solution address the root issue(s)? When relevant, please provide data-backed (i.e. citation, research, reports, statistics, etc) arguments to support your claim.*

Since our project is very new, we have not yet any evidence for the impact of our work.

Using our layered data integration approach in a “custom impact field generator” shows users the personalized/customized context sensitive information in an easily digestible way – a customized map with hot spots and cold zones related to the problem specified.

We know that gathering relevant information and managing data assets is more and more an increasing organizational problem. Although technology improves, day 2 day procedures are often well behind. With our data integration solution, we can support people in industry and research in a variety of fields to rethink the way how they exchange and share and collaborate on data assets.   
 Further reducing access barriers to relevant information is our social impact, and reduced data provisioning cost is the economic benefit of our approach. Assuming, that all major open data platforms would collaborate in our planed initiative, we can be sure to see a global movement, using and contributing to open data collections in a variety of domains.  
 Embedding the software module into mobile applications and SaaS solutions leads to a wide range for adoption. (191)

\* Please explain the following (Max 250 words) for Innovativeness:

***How does your project offer a novel use of graph or approach to the problem?*** *In your response, address the following: Did you frame a complicated problem in a new way? Did you come up with a creative way to overcome the challenges? When relevant, please provide data-backed (i.e. citation, research, reports, statistics, etc) arguments to support your claim.*

Instead of thinking about ETL tools, big data platforms, or data pipelines, we treat contextualized information, managed in linked data pods, as the essence of a global information architecture. Personal responsibility and comprehensive data integration, not on the tool level, but rather on the data model level is our approach – and this is not yet a standard, but to be honest, it is not new and we didn’t invent it ☺  
The graph based approach is already inherent in the Linked Open Data cloud (LOD) - but the property graph as used in TigerGraph is of a different structure. While the linked open data cloud is focused on direct interlinking of data items, the TigerGraph platform offers scalable graph processing and graph analysis capabilities, which is not available in most of existing linked data systems. Our approach brings both worlds - RDF graphs and Property Graphs - closer to each other. We do data integration on the graph data model using a small set of fundamental linking and indexing techniques.

There is not yet an established and simple way to blend data assets. We develop and study a set of fundamental graph operations, such as layer projection, vertex attribute integration, node attribute propagation.   
 To our knowledge, such a set of minimal operations which allows us to blend and 'morphe' data models does not yet exist.   
 Our ambition is to practice a new kind of graph data management procedures. With partners we want to establish a global sustainability science data graph.

(250)

\* Please explain the following (Max 250 words) for Ambitiousness:

***How does your project push the boundaries of, scale and scope of graph to solve real world problems?*** *In your response, address the following: What is your graph schema size (entity types, relationship types, domains, etc? What’s the scale of your graph solution and the functional features you’ve added? When relevant, please provide data-backed (i.e. citation, research, reports, statistics, etc) arguments to support your claim.*

Datasets from many real-world projects are rarely combinable. This is because typically a lot of domain knowledge is needed to bring data assets from different context together.   
 Because our approach is focused on simplification we would be able to bring the full OSM dataset, the Full Wikidata dataset and many others into just one layered graph. The problem is the scale of the resulting graph. We solve this problem by using regional contexts as partitioning mechanism. Handling small partitions of all aspects is still possible on a developer machine, or a free cloud instance – but since our approach uses TigerGraph’s property graph, it scales with the capabilities of TigerGraph.   
 Our goal is to see the most important public data assets combined in one large graph, so that researchers of multiple domains can do their own projections in place.

The question we address today is: How sustainable can I life in a specific region? This is a very specidic question, and also a very ambitious goal, because a lot of information needed to answer this question is not yet available.   
 Hence, besides the graph integration procedures, we also need collaboration with the open data initiatives, such as Wikidata, OpenStreetmap and many others on a global scale, and not just owned by one of the big players. (215)

\* Please explain the following (Max 250 words) for Applicability:

***How will your project be adopted and applied by other organizations or industries?*** *In your response, address the following: How easy will it be to put your solution into real-world use? How many industries could adopt your solution? What is the size of this market or industry that could adopt your solution? When relevant, please provide data-backed (i.e. citation, research, reports, statistics, etc) arguments to support your claim.*

We started with a specific use case: real estate industry. The question is: How sustainable can I life in a specific region? This works as a simple example anybody can understand. We will then search for collaborators in the context of sustainable merchants, smart cities, sustainable supply chains, and most importantly, the sustainability industry which is currently emerging. A detailed market analysis is work in progress. (66)